DOCUMENT RESUME

ED 052 244 TM 000 650

AUTHOR	Womer, Frank B.
TITLE	Research Issues Arising from the National Assessment of Educational Progress.
PUB DATE	Feb 71
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Academic Achievement, *Data Analysis, Demonstration Programs, *Educational Objectives, Educational Research, Evaluation Methods, Models, *National Surveys, Performance, *Sampling, Student Testing, Symposia, Test Construction, Testing Problems, Test Reliability
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

This symposium deals with recent issues in the development of the National Assessment model. General goals are outlined and the following topics are discussed: "Objectives and Exercises" (Jack C. Merwin); "Sampling" (A. Finkner); and "Data Analysis" (John Milholland). (CK)

Frank Womer, Staff Dir, NAEP, Ann Arbor, Mich Feb 71

GENERAL GOALS

This portion of the symposium will raise research issues related to the development of the National Assessment model, to the purposes and structure of the project. Specific issues to be discussed are:

1. Should the assessment project be (remain) national only or should it be extended to other smaller legal jurisdictions?

2. How closely should National Assessment reflect political desires for "comparative" information versus its original goal of providing descriptive information?

3. Should National Assessment attempt to serve as a model for state and/or local assessments?

4. Should National Assessment remain as an information-gathering, census-like project or should it more actively seek to relate its results to various in-put data from schools?

5. Should National Assessment continue to report group results only or should it change its goals and methods to enable individual pupil reports to be made?

6. Should National Assessment broaden its reporting categories, should it reduce its reporting categories, or should it keep them as they are?

7. Should National Assessment confine its data collection to its own purposes or should it allow other significant studies to be piggybacked onto it?

8. Should National Assessment continue to develop its own objectives for each subject matter area and the allied materials or should it seek to use existing statements and/or materials?

9. Should National Assessment continue to assess subject matter areas primarily or should it try to look at the general goals of education without subject matter constrictions?

10. Should National Assessment attempt to develop an index or indices of educational achievement, a GroNK (Gross National Knowledge)?

11. Should National Assessment set up a data bank to make its data available to other researchers, as does Project Talent?

12. How can one reconcile the practitioner's demands for "instant" information, for "pat" answers, for dramatic headlines with the researchers hesitancy to use untried materials or to release results prematurely?

13. Should National Assessment expect to research questions that arise in relation to its own operations (as it formerly did), should National Assessment "armchair" such questions (as it is doing now), or should it turn to the educational research community for answers (as it has done occasionally)?

14. Should National Assessment attempt to interpret its results or only report the facts?

15. To what extent will institutionalization be a problem with National Assessment?

Jack C. Merwin, Dean of School of Ed
Univ of Minnesota
Minneapolis, MN

Feb '71

OBJECTIVES AND EXERCISES

Early in the history of the National Assessment Project three decisions with important implications for instrumentation were made. The first of these was the decision that the assessment would be based on educational objectives; objectives to be those being taught for in the schools, and accepted as authentic to the disciplines by scholars in the relevant field, and considered by the lay public as desirable goals for American youth. A second decision was that reporting would be on the basis of responses to individual exercises rather than in terms of scores across exercises as is common practice in testing. A third decision, not unrelated to these other two, was that exercises were to be developed which would help describe what practically all students of a given age level can do and what the most advanced of an age level could do as well as what the average could do. Two basic considerations for the development of instruments grow out of these decisions.

It was obvious from the start that if individual exercises were to be exhibited in the reporting and such reporting was to be meaningful to nonprofessionals and professionals alike, they would require a degree of face validity seldom imposed on the individual items of achievement

tests. Each exercise would have to be able to "stand alone" with the data on that exercise and carry some meaningful information about what American youth can do. While this would be similar to the reporting done by survey opinion pollers it was a demand for content or face validity unlike that previously required in the area of achievement assessment.

The second requirement for instrumentation that grew out of the early decisions is what has been labeled "directionality". Objectives were to serve as the basis of instrumentation covering exercises from both the cognitive domain and the affective domain. Thus, it was necessary that there be a means of classifying responses using a dichotomy, an ordered classification scheme or some dimension of desirability of response. In other words, if the schools are teaching for something, then some responses ought to be preferable to other responses that might be given. Exercises which do not have such "directionality" and became simply surveys of information would not be satisfactory for the assessment project.

The approach to instrumentation was to first contract with test development agencies to work with professionals on the development of objectives. These objectives were then reviewed by lay people and a set which in general met criteria mentioned earlier were derived. Contracts were then let for the development of exercises to tap those objectives. Early in the game some had naively believed that to assess what groups can do, as opposed to testing what individuals can do, would be a simple matter of pulling existing tests or items off of the

shelf and administering them to a group of people. It early became obvious that this would not provide the instrumentation needed to meet the criteria noted above. The contractors produced and submitted to the staff of the project some ten thousand unique items which, with overlap, numbered twenty thousand items proposed for use at the four age levels of 9, 13, 17 and adult. On reviewing these exercises several characteristics were quickly identified. First, it was quite clear that those who had prepared the exercises had not taken advantage of the freedom to be creative in approaches to developing the exercises since many restrictions involved in developing reliable measures of individuals were not present. The second observation, later documented through the results of a study, indicated that the exercise developers had little conception of the type of exercise needed to describe what 90 per cent or more of a given age group could accomplish. Based on a rough and ready sample, in one subject area the easiest of the items to be proposed for use to describe what practically all students at the nine-year old level could accomplish was answered correctly by just 55 per cent of the students involved. The p values for items in all areas that had been developed as 90 per cent items ranged down to p values of 0-4. A third observation was that the high degree of face validity called for by the reporting scheme had not been built into a large majority of the exercises produced.

A series of reviews, tryouts, and studies were undertaken by the National Assessment staff in an attempt to identify specific problems and find ways to deal with the exercises at hand and the deficiencies

that had been suspected or identified in them. Reviews by subject matter specialists not involved in the development of the exercises were conducted for the purpose of examining the content validity of each exercise in regard to the objective it was to tap, the appropriateness of keying and identification of any ambiguities or other potential difficulties with the exercises. It was ultimately necessary to carryout both a mail review and individual working review conferences where subject matter experts could interact in an attempt to rework exercises into acceptable form. Reviews with lay people were conducted in an attempt to identify the extent to which the exercises might be meaningful if used in reporting to nonprofessionals and also to identify any exercises with aspects that would be considered objectionable to be asking American youth to deal with. Studies to determine the extent to which we indeed had "easy" enough exercises to describe what practically all of the youth at a given age could do, the feasibility of administration of exercises as proposed by the contractor, the need to augment presentation of exercises to the student so that errors in the communication of the task were minimized as a source for a wrong answer or no response and a number of other studies were undertaken.

Let me turn then to what has been learned so far and some of the problems that remain unsolved and in some cases not clearly delineated.

It soon became apparent that to meet the criteria set up for instruments to assess what large groups can do, that few guidelines were forthcoming from the conventional testing approach for individual students. Some help was to be found in the experience of opinion poll people in accumulating responses from groups but much unplowed territory

remained in the area of instrument development for assessing the achievement of groups. A second matter which became quite obvious was that it is very difficult to get people normally attuned to constructing test items which are to become merged with other items to form a test for individuals with concern for efficiency in getting reliable measures of individuals to exercise creativity when not faced with restrictions that individual testing imposes. For instance, it was made perfectly clear that an exercise that might take up to as much as thrity minutes of performance time, if indeed it would provide valuable information on the achievement of one of the objectives, would be permissible. As I will note a little later, this had to be temporized by practical considerations in a number of instances. On the other hand there were not a large number of exercises developed which met the criteria and involved creativity to which such practical limitations were imposed. A third major problem was the limited amount of time, even with an unmatched design and using fourteen packages of exercises at an age level (i.e., it took fourteen students, each responding to a different package to get one response to the set of items used at an age level) the most appropriate compromise between numbers of objectives covered and the density with which an objective is covered is yet far from adequately solved.

In spite of the above difficulties, experience has been gained through the administration of exercises on a national scale in year one and into year two in areas that could be most readily shaped up for use.

Major Questions Currently Under Examination

1. Since it is seemingly inefficient to simply contract for X number of exercises to tap an objective, might a prototype approach based on the work of Hively and Osburne be more efficient? That is, might the operation better call for a two-step operation? The first would be the development of a prototype with identification of the "shell" to be used to generate other items and the variables and the limits on variables to be allowed in generating items based on that prototype.

2. How can the results and feedback of experience with the results from the first reporting best serve as a basis for identifying difficulties and working toward their resolution?

3. Since the procedure is to report some exercises and not report others there will be the need to develop exercises to replace those reported for the next go-around in that particular subject area. The question then is what different procedures, if any, are needed for developing replacement exercises over and above those for developing the original exercises?

4. How can the overall efficiency of producing the instrumentation be increased, particularly in regard to the review and revision for updating of objectives, the development review and tryout of exercises, and the process of selecting from among exercises and acceptable exercise pool for packaging so as to most effectively cover the objectives under consideration?

A. Finkner, Res Tri. Inst
Res. Triangle Park, N. C.
Feb 71

SAMPLING

From the survey sampling point of view, there are two primary requirements to be placed on the sample. First, every element of the population must have some known probability of being selected and secondly, the procedure must be operationally feasible in the field. Of course, the general principle of modern probability sampling must be adhered to insofar as possible. This principle states that the design will provide maximum information at a given cost. In addition, however, the unique character of the National Assessment of Educational Progress placed somewhat unusual constraints upon the sampling design. Some of these constraints are given below:

(1) Only one package of exercises can be administered to one child in a given school.

(2) In order to cover an objective, many packages of exercises had to be developed.

(3) Each package contained exercises from all three objectives in the Year 01 assessment with various levels of difficulty for each exercise.

(4) To date there has been little consideration of combining scores on selected items into an index; hence, the actual sample size is a number of students administered a particular package, not the total number of students of a given age participating.

(5) Procedures for the calculation of sampling error estimates must be provided. Control of SES makes this difficult.

(6) At least two schools must be represented in each PSU, i.e., no one school should be administered more than 10 packages, if possible.

(7) Mode of administration, i.e., tape administration reduces the options to one package per sitting. Such a constraint tends to increase the clustering effect on the variance.

In addition to these constraints, a number of attributes of the sampling design were considered desirable. Some of these were:

(1) Estimates were desired by sub-population for some 256 sub-populations: four regions by four types of community by sex by two levels of socio-educational status and by four age groups. Eventual categories for reporting could not always be defined prior to sampling; in some cases, if a definition is possible, field identification prior to sampling is not.

(2) At least three kinds of comparisons were desired.

(a) For example, comparisons of certain sub-populations on the basis of the responses to specified science exercises in Year 01.

(b) Comparison of responses in Year 01 with responses in Year 04 for a given sub-population.

(c) Comparisons of differences between sub-populations between Year 01 and Year 04.

(3) The costs of selecting the sample and for administering the packages in the field should be held to a minimum. In this connection, what are the advantages of household vs. mixed frames for all age groups?

(4) Since estimates by states were not to be provided, no attempt was made to include all states in Year 01. For public relations reasons, it was deemed desirable to include all 50 states in Year 02.

These constraints and desired attributes lead to a number of considerations. Some of these are listed as follows:

(1) Age groups 9 and 13 are essentially found in schools. The group between the ages of 26 and 35 are essentially found out of school. Although age 17 is mainly in school, a sizeable portion of that age group is no longer in the secondary school system for one reason or another. What kind of a sampling frame or multiple sampling frame needs to be devised in order to give every person in each of these age groups their proper chance for selection?

(2) If a student cannot provide information on his socio-educational status, how should this information be obtained?

(3) Since some students sampled in schools will be clustered, their performances will be positively correlated. How much should the sample size be increased to compensate for this lack of independence?

(4) Should the sample be rotated so as to avoid, insofar as possible, placing a burden on any school system? How can rotation be implemented to utilize the correlation between years and inter-year comparisons? How do we maintain comparability in the definitions of the sub-populations over time?

(5) If a multi-stage sampling design is deemed to provide the most information per unit cost, how should the primary sampling units be structured?

(6) What size of cluster of students in the schools is most efficient?

(7) Should small sub-populations which are of particular interest be oversampled to increase the precision of the assessment for such sub-populations?

(8) Should the present NAEP sample design be adopted by states if they should desire to conduct similar assessments?

(9) What information on sampling should be made available to the public, i.e., frames, programs for estimating "p" values and sampling errors, and sampling efficiencies?

(10) What are the categories of non-response and what is its effect on the estimates?

(11) Can more effective stratification be devised?

(12) To what extent should we and can we link Census data to the data we collect?

(13) How do we sample 17-year-olds out of school?

Dr John Milholland,
Univ of Mich, Ann Arbor,
Mich Feb 71

DATA ANALYSIS

There are two reasons why I feel somewhat hesitant about speaking on the topic of problems of data analysis in National Assessment. One is that I was a staff member for only a short time and really did not have the opportunity to try to unravel the various knots with which I was confronted. Most of my time was spent in the day-to-day decisions that were involved in carrying out policies and procedures that were already underway.

The second reason is that the ground work for data analysis and the solution of the really intricate problems of dealing with the results of national assessment have been in the hands of an extremely competent committee called the Analysis Advisory Committee. Its membership consists of John Tukey, chairman, Robert Abelson, William Coffman, Wayne Holtzman, Lyle Jones, Fred Mosteller, and Ralph Tyler. Lee Cronbach was formerly a member. It would be my judgment that if there are any problems of data analysis that they can't handle it would be foolish for me to try even to understand them. As I have observed them operate I have found that they not only consider the major issues of taking care of the data but are also willing to penetrate deeply into the results themselves so that they are thoroughly familiar with just what is going on. I think National Assessment has been extremely lucky in having a committee of this eminence exhibit the amount of dedication that they have done. Among the many problems they have faced have been the determination of appropriate standard errors for subgroup differences; developing a basis for classifying schools and localities in accordance with educationally relevant characteristics so that differences in educational attainments associated with different types of communities might be studied; and the derivation of an index of performance characteristics for the exercises which to a certain extent measures the advantage-disadvantage dimension of educational accomplishment. These were

exciting, difficult, and interesting problems and a discussion of them by a member of the Committee at this point might have been preferable to what is actually being done. We are so jealous of their time, however, that we didn't even approach any of them about it. I shall not attempt to go over the ground that they have covered (although I might be able to respond to some questions about the items just mentioned), but will simply remark upon what appear to me to be some rather persistent problems of data handling that National Assessment is faced with. Perhaps there are no real solutions but it is well to know what the problems are.

There is a possibility that one problem may arise out of the way National Assessment is being done. Usually in a survey all respondents take the entire assessment instrument. In Project Talent, for example, high school students spent the better part of two days taking the test battery. One of the features of National Assessment which I am sure is at least partly responsible for the high level of cooperation among school systems is that no pupil invests more than about an hour of his time with the assessment. The exercises are packaged in separate booklets, each of which takes less than an hour to administer. Each survey involves somewhere between ten and fifteen of these packages. Since exercises, as well as persons, are sampled, any combining of exercises across packages will introduce a "package" effect. This could happen if an attempt was made to add together the performance of respondents to exercises that contributed to one objective but appeared in different packages. The following situation might occur.

Package:	A	B	C	X	Y	Z
Objective I (No. of Exercises)	3	5	7	9	11	13
Objective II (No. of Exercises)	13	11	9	7	5	3

Packages X, Y and Z are heavily weighted with Objective I; A, B, and C with Objective II. If accidents of sampling have resulted in higher quality respondents for one of these sets, the corresponding objective will appear to have been better mastered.

One solution for this problem is to make comparisons that involve only single exercises; another is to balance across packages the number of exercises contributing to any objectives.

The first option would provide for regional, type of community, and so forth comparisons being made for one exercise at a time and it might be expected that patterns, not necessarily levels, of ability would be constant. This could be extended to comparisons between different exercises in the same package but the boundaries of the package would provide a barrier which could not be crossed.

The balancing of objectives across packages probably has been achieved to a considerable extent, although ascertaining the exact degree of success in this is just one of those things I never got around to do. Although it would be my judgment that the package affect is generally of small moment, it is something that one should be aware of particularly if some new bases of combining exercises are instituted.

Dr. Finkner has already mentioned that eventual categories for reporting could not always be defined prior to sampling and that in some cases even if a definition was constructed field identification prior to sampling was not possible. Although a number of variables were brought under control by stratification it was not possible to control all potential categories that we might be interested in. It was expected, however, that there would be a good enough distribution for some in the sample to make comparisons reasonable. For example, it was anticipated that the sex ratio would be satisfactorily distributed with respect to the major factors in educational attainment.

This is apparently not true for skin color, however, which tends to get confounded with region, type of community, and parent educational level. When the sampling plan was originally set up it was not intended that comparisons on this basis be made. Later it seemed important to do it, and in our first reports there are and will be some comparisons of Blacks with non-Blacks. In later reports the break down will be Black, White, and Other. Members of certain subgroups of the "Other" category are already asking that some separation of results for them be instituted. This kind of adventure should be entered upon only with great caution since some of the subgroups are not well distributed throughout the country, and some comparisons might partake more of affects generated by extraneous variables than by the defining characteristic of the subgroup itself. Our problem will be how to analyze and present the data in such a way that false inferences cannot easily be drawn from them.

Some of the confounding just referred to leads to an attempt to make some statistical adjustments. For example, if we found many more 13-year-old boys than girls in some region, and this region turned out to show superiority in mathematical and mechanical exercises, someone would be likely to point out that it was a sex, not region, effect. We can balance this, of course, with some safety, but it becomes more tricky when we deal with type of community interacting with educational level of parent, for example. Would it really be true that inner city children with well educated parents would exhibit the same performance if they moved to the affluent suburbs? Would Black chilren in the Southeast perform the same if they moved to Minnesota?

There will be a number of assumptions and constraints that will inhere in any procedure of statistical adjustment and it is essential that they be known and explained. The Analysis Advisory Committee is working on the problem of statistical adjustment and I am sure that if anything can be done with it they will be able to do it.

A special problem has arisen in connection with the assessment of writing. In the various subject matter areas to be covered in National Assessment the sequence of exercise development has been to gain agreement upon objectives, and then, in line with them, to identify some knowledges and abilities that the clients of our educational system should, with certain probabilities, possess. In writing, for example, it is reasonable to expect that our populace should be able to fill out an application blank, order something by mail, and address an envelope properly. Standards of acceptable performance for these activities can be set.

When it comes to unstructured, or, euphemistically, creative, writing the picture is different. No one seems to want to write a model essay and say, "85 (or 50, or 15) percent of 13-year-olds should write this well." As a result, responses to exercises of this sort were scored by committees of readers who set up their scales on the basis of the writing produced by the respondents. This was indeed a norm-referenced scoring, and certainly will lead to problems of analysis when we go to make an assessment of change the next time writing is measured. A preview has been furnished us by our attempt to obtain a common scale for some exercises that were administered at two age levels. In our November report we did try to show comparisons of 13s with 17s on one exercise. There's an Appendix D which describes the procedures we followed, and it furnishes an example of problems of data analysis.

National Assessment's avowed intent to report the facts as they find them without interpretation of course opens the door to interpretation by anyone who wants to do it. One assumption which may lie buried too deeply in our operations is that the results present a true picture of what young Americans know and can do. This, however, will be so only if the exercises somehow appeal equally to all respondents. We have already been challenged to take into account the role that motivation might play in performance on our exercises. Certainly the experiences of the past few years have alerted us to

the fact that many Black Americans simply do not put their hearts into striving to do well on the usual kind of assessment devices. Other subgroups, perhaps most adults, may have similar reactions. I think one of the major problems National Assessment is going to have to face is the one of gaining some assurance that its exercises are attacked by various groups with somewhere near the same level of effort.

A related problem that is caused by a reaction against testing generally, which is exhibited to greater degrees by some members of our population than by others. If refusals for this reason run higher in some parts of our sample than in others a response bias that is difficult to evaluate may appear.

In the last year or so a number of national organizations interested in measurement and testing have been faced with proposals to institute a moratorium on testing of minority groups until cultural biases can be removed from the tests. I think there are better solutions to the problems of bias than abandoning testing altogether but if this extreme position is actually taken and becomes fairly effective National Assessment could no longer claim to be a survey of educational attainments in the country. One course of action that might be followed would be to subdivide the population into various groups and try to develop appropriate assessment instruments for each. If this were to happen the problems of data analysis would indeed be magnified, and I think I shall retire and contemplate them.